Projet 6 : Avis Restau

Challenge : Améliorez le produit IA de votre start-up

https://github.com/blanchonnicolas/IA_Project6_Openclassrooms_IAStart_Up

Agenda

01

Traitement des données textuelles

Contexte et objectifs
Jeu de données
Nettoyage des données
Pondération des mots
Etude et Sauvegarde des modèles de classification

02

Classer de nouvelles données collectées via API YELP

Authentification API
Structure JSON GraphQL
Pipeline Traitement du texte
Chargement des modèles
Classification des nouvelles données

03

Représentation des résultats sur une PAGE WEB

Création des variables RFM
Création d'autres variables
Visualisation en Composantes principales
Analyse des segmentations utilisant KMeans, ACH et DBScan

04

Traitement des données Images

SIFT (Descripteurs)
Segmentations KMeans
Réduction de dimension (ACP et T-SNE)
Analyse de similarité

Descriptif du contexte projet

Détecter les insatisfactions

Analyser les commentaires postés sur la plateforme.

Extraire les sujets d’insatisfactions les plus communs.

Labelliser les photos

Identifier les photos relatives à la nourriture, les boissons, le décor intérieur ou extérieur du restaurant, par l’utilisation d’un classifieur non entrainé.

Segmentez des clients d'un site e-commerce

Proposer une étude sur la faisabilité
- de détection des sujets d'insatisfaction
- la labellisation automatique des photos.

Déterminer les sujet d'insatisfaction
Labelliser automatiquement des images
Page Web permettant d'utiliser les modèles de classification

Fichiers JSON (reviews et photos)
Dossier Photos
API YELP

Traitement NLP (Texte) et SIFT + CNN (Images)
Réduction de dimension
Classification non supervisée
Visualisation

Back to AGENDA

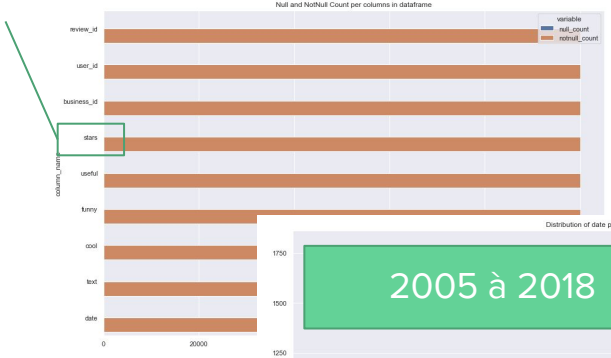
Détecter les sujets insatisfactions à partir de données textuelles: *Review*

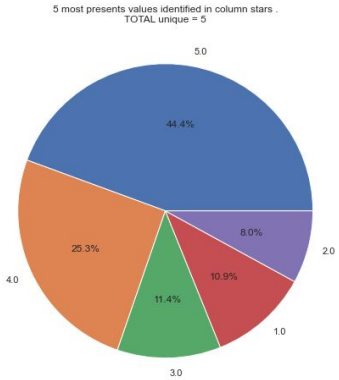
Présentation du jeu de données

Compréhension du jeu de données

yelp_academic_dataset_review.json
- 4.97Go
- n lignes importées = 200000
- Filtrage sur les satisfaction faibles

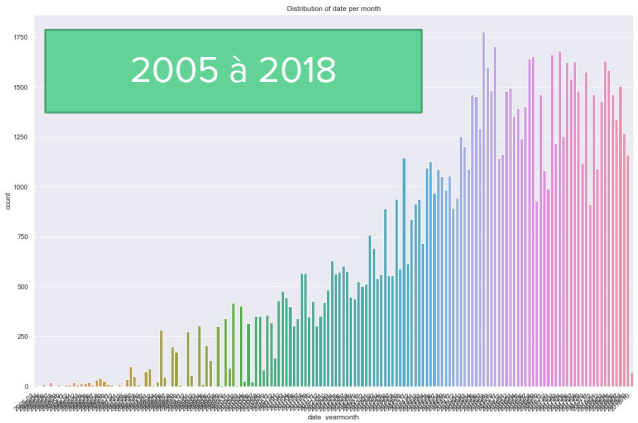
Pas de valeurs nulles / vides





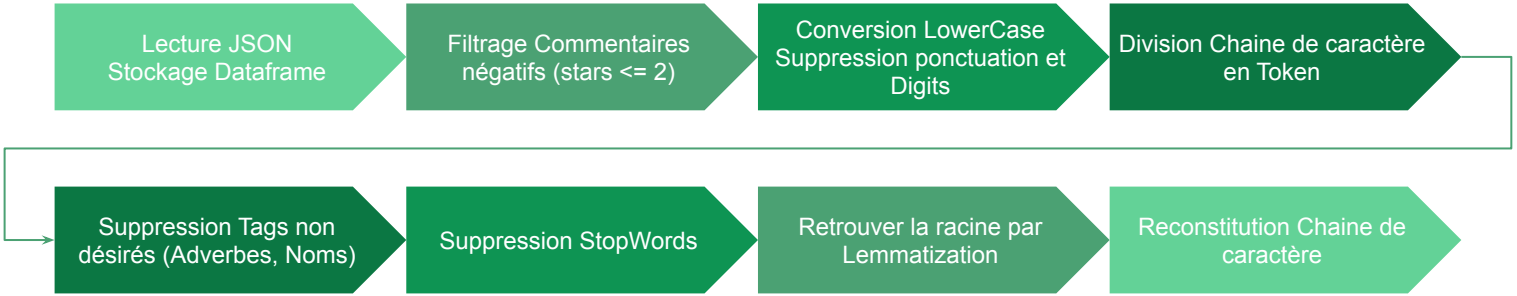
	stars
5.0	44392
4.0	25337
3.0	11362
1.0	10921
2.0	7988

Filtrage Notes de 1 à 2



Analyse des sujets d’insatisfactions

Préparation des données

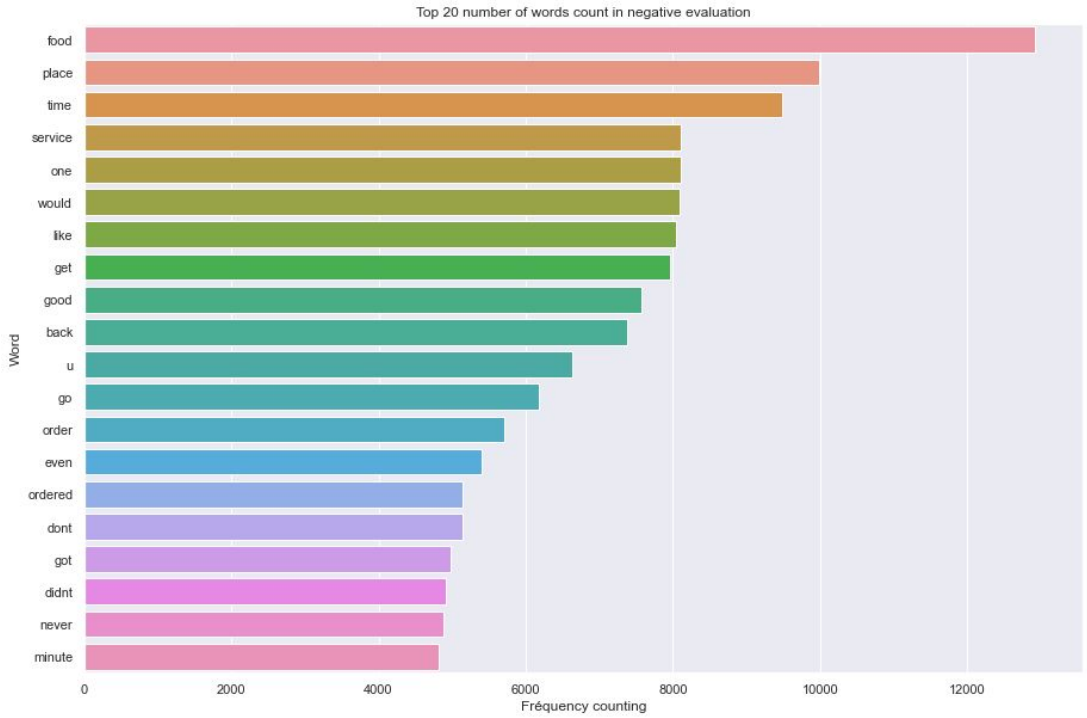


Note: Etape ajoutée après analyse des Tags

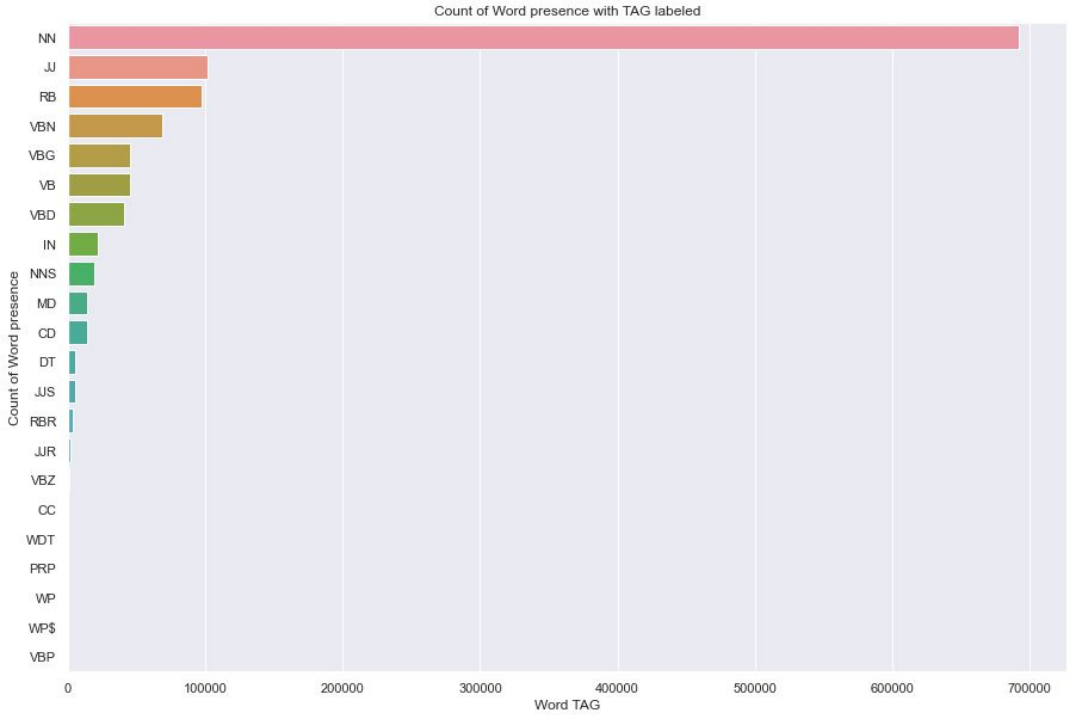
Nous construirons un “pipeline” afin de préparer de futurs jeux de données, et assurer la reproductibilité de la préparation.

Comptage de la fréquence des mots

Nous notons ici que les mots les plus fréquents, appartenant au commentaires associés aux évaluations négatives n'appartiennent pas à la même famille (Noms, Adjectifs, Adverbes, ...)



Etude de la nature des mots



Nous voyons ici qu'une forte proportion des mots sont de Nature = Noms (NN).

L'étude approfondie de chaque TAG nous permettra de sélectionner les tags les plus utiles dans le cadre de notre analyse.

Représentation visuelles en Nuage de mots

Tags excluded from Wordcloud are tag_type_to_eliminate = ['RB','RBR','MD','CD']

Nous représentons ici les 50 mots les plus fréquents dans le corpus.

Plus leur nombre d'apparition est fréquent, et plus la taille est imposante.

A ce stade, il reste complexe de déterminer de vrais sujets d'insatisfaction.

Pondération des mots

Pondération des Mots

Nous évaluerons l'importance des mots, par leur fréquence, ainsi que par leur singularité.

Pour se faire, nous utilisons la métrique tf-idf (Term-Frequency - Inverse Document Frequency) qui permettra de pondérer le nombre d'apparition du mot par rapport à sa rareté.

Nous obtenons une matrice tfidf, correspondant à une représentation vectorielle des mots.

Nous sauvegardons le modèle pour une future réutilisation.

Created 38038 X 1546 TF-IDF-normalized document-term matrix

	0	1	2	3	4	5	6	7	8	9	...	1536	1537	
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.21
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.00
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.00
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.154359	0.0	0.0	...	0.0	0.0	0.00
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.00

TfidfVectorizer(min_df=0.005,max_df=0.8)

Classification non-supervisées

3 Algorithmes de classifications

LDA (x2)

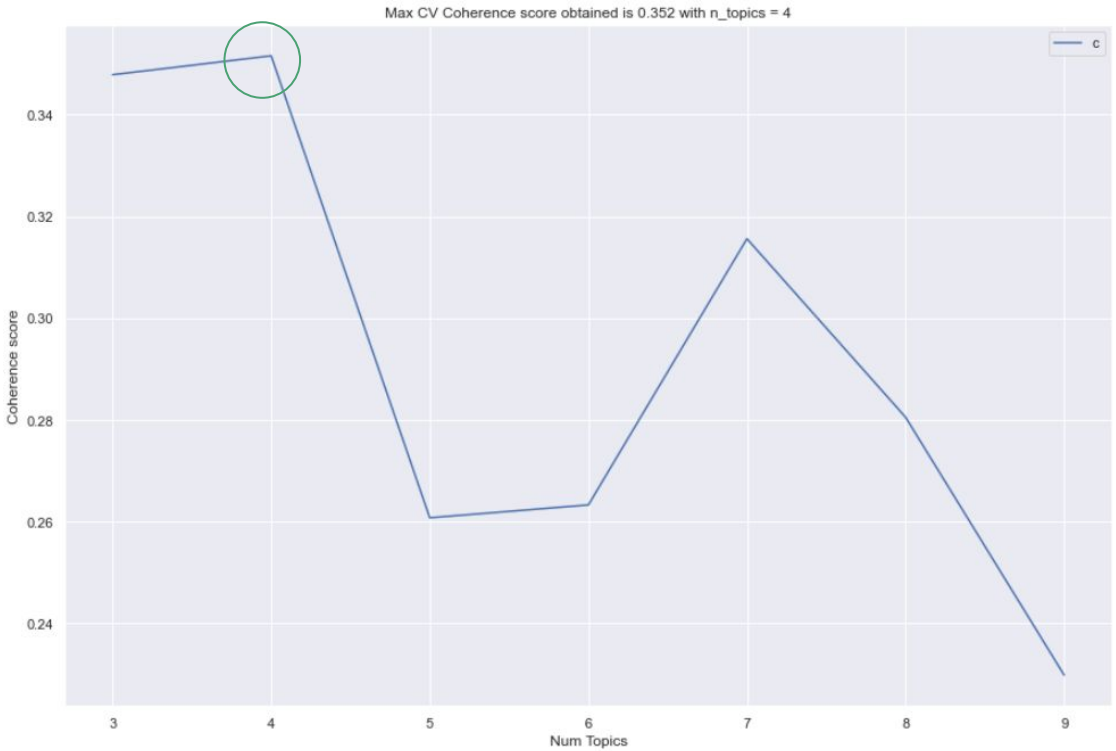
Modèle probabiliste génératif qui permet de décrire des collections de documents de texte.
Il nous permettra d'étudier des structures thématiques cachées dans le texte, et ainsi déterminer des thèmes.
Nous utiliserons pour cet algorithme 2 librairies (sklearn et Gensim)

NMF

Non-Negative Matrix factorization
La NMF est une technique de réduction de dimension adaptée aux matrices creuses, par exemple des occurrences ou dénombrements de mots.

Nous essayons ces algorithmes avec un nombre non optimisé de Topics (=5), ceci afin d'obtenir un premier aperçu des classifications proposées.

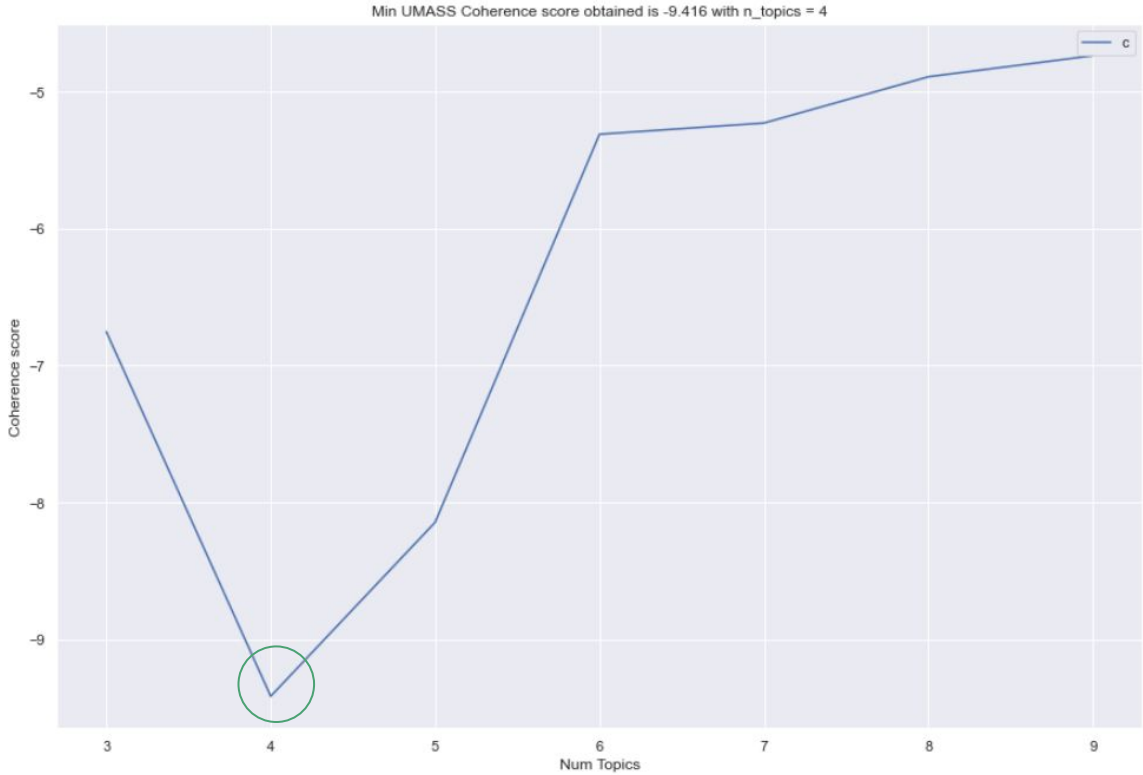
Recherche d’optimisation des hyper-paramètres



L’utilisation de la librairie GENSIM LDA nous permet de rechercher le nombre optimal de Thème, par la mesure du score de cohérence C_V.

Le choix optimal de cette mesure s'apparente à la méthode du coude, et nous privilégions ici Topic = 4.

Recherche d’optimisation des hyper-paramètres

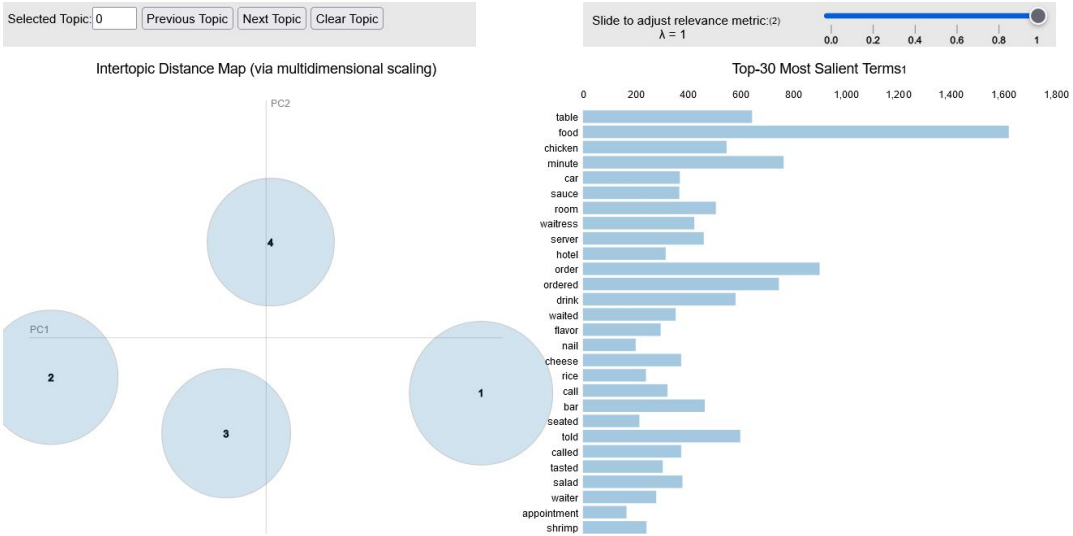


Nous recherchons ici le score de cohérence U-MASS optimale, en prenant en compte l’apparition simultanée des bi-grammes: Fréquence à laquelle les deux mots (Wi, Wj) ont été vus ensemble dans le corpus.

Représentation visuelle des Thèmes

La librairie pyLDAvis permet de faciliter l'interprétation des thèmes à l'aide d'une représentation graphique sur 2 composantes principales.

Nous pouvons ainsi tirer des conclusions sur l'importance de certains mots pour certains thèmes.



	stars	text	0	1	2	3	Topic
0	1.0	i am a long term frequent customer of this est...	0.048141	0.851809	0.050754	0.049296	1
1	2.0	i at least have to give this restaurant two st...	0.048930	0.045696	0.859337	0.046037	2
2	2.0	straight to the point its cheap it tastes and ...	0.889380	0.039621	0.035980	0.035019	0
3	2.0	never again this is a so called restaurant tha...	0.043753	0.869885	0.042910	0.043451	1
4	1.0	if you want to pay for everything a la carte t...	0.526079	0.036791	0.400338	0.036792	0
...	...	...	...	...	...	...	...
38033	2.0	the whole place has just a few souvenir stores...	0.060190	0.820500	0.059149	0.060161	1
38034	2.0	sorry spiros i didnt want to say anything bad ...	0.037490	0.037322	0.888587	0.036602	2
38035	1.0	this place sucks we ordered food for delivery...	0.805584	0.040710	0.041809	0.111897	0
38036	1.0	i spoke with keri about a month ago on the pho...	0.032661	0.033263	0.033519	0.900558	3
38037	1.0	the hotel checked me out early without my perm...	0.042531	0.043087	0.043902	0.870479	3

- Topic 1 : Nourriture, Aliments et Goûts
- Topic 2 : Réservation
- Topic 3 : Bar et boissons
- Topic 4 : Temps d'attente et Service

Back to AGENDA

Ajouter de nouveaux commentaires, et les classer dans le thèmes prédéfinis

API GraphQL & Process de classification

Authenticate and Run the Query on GraphQL API

Store API return in files (stores in JSON and converted in CSV), Load in dataframe and process text preparation pipeline

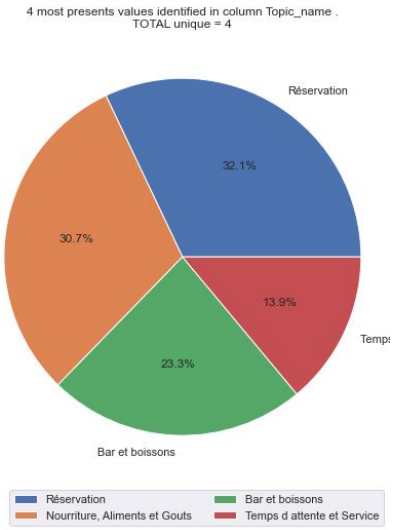
Run LDA and GENSIM classifiers

```
Query = search(location: "paris", limit: 50 offset: ++) {
        business {
                name
                reviews {
                        rating
                        text
                        }
                }
        }
```

Chargement et réutilisation des modèles:

- tfidf = load('./Models/tfidfvectorizer.joblib')
- lda_tfidf_api = load('./Models/lda_tfidf.joblib')
- tfidf_gensim = load('./Models/tfidf_gensim.joblib')
- lda_gensim_api = load('./Models/lda_gensim.joblib')

Représentation visuelle des Thèmes



text	name	0	1	2	3	Topic	Topic_name
ck and forth trying to figure out wha...	Le Comptoir de la Gastronomie	0.063620	0.808858	0.063878	0.063644	1	Réservation
ourist trap restaurant this restaura...	Le Comptoir de la Gastronomie	0.079374	0.762068	0.079165	0.079393	1	Réservation
ppointed this time around as hubby wanted ...	Berthillon	0.063654	0.310786	0.065063	0.560497	3	Temps d attente et Service
place is permanently closed please update...	La Coincidence	0.068485	0.070398	0.069626	0.791491	3	Temps d attente et Service
restaurant but very packed and very sma...	Eggs & Co	0.067211	0.803567	0.066689	0.062533	1	Réservation
...	...	...	...	...	...	...	...
s in a lovely vacation in the park by ...	The Pavilion Cafe	0.065000	0.805708	0.063501	0.065792	1	Réservation
l desert were excellent appetizer o...	Locanda Locatelli	0.783298	0.071832	0.074606	0.070264	0	Nourriture, Aliments et Gouts
ht be the best burrito in london wow th...	Daddy Donkey Mexican Grill	0.824170	0.060227	0.057989	0.057614	0	Nourriture, Aliments et Gouts
lescribe my experience here as very un...	Sen Viet	0.772822	0.078068	0.073942	0.075168	0	Nourriture, Aliments et Gouts
ecked in at 4pm on a sunday for our reservat...	Roast	0.063575	0.065103	0.805592	0.065729	2	Bar et boissons

Nous mesurons la proportion des avis, sur chacun des thèmes obtenus. Ensuite, nous représentons visuellement l'association mots / thèmes afin d'interpréter les résultats.

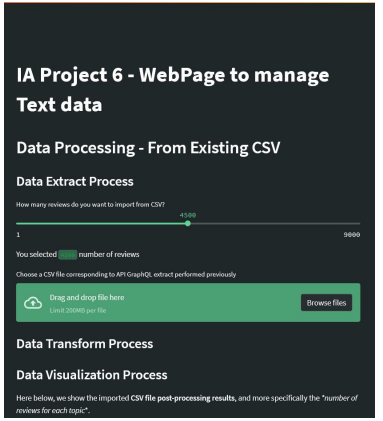
Back to AGENDA

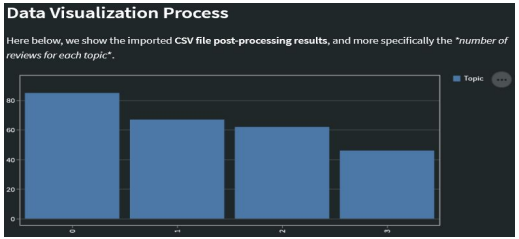
Représentation des résultats sur une PAGE WEB

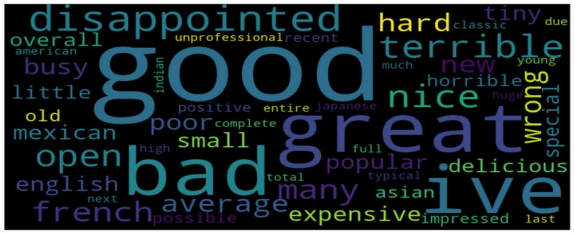
Traitement des données reçues par l'API

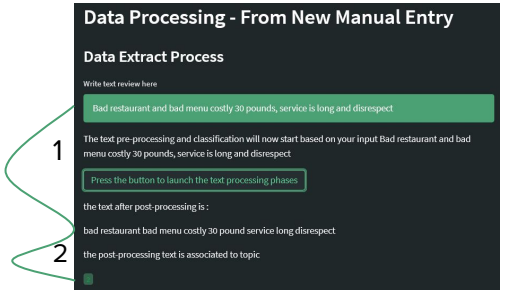
Utilisation de Streamlit pour générer une page HTML interactive.

1. Choix du fichier CSV à importer (voir données de l'API)
2. Traitement / Nettoyage Texte via pipeline
3. Classification des sujets + WordCloud (réutilisation de modèle via "joblib")









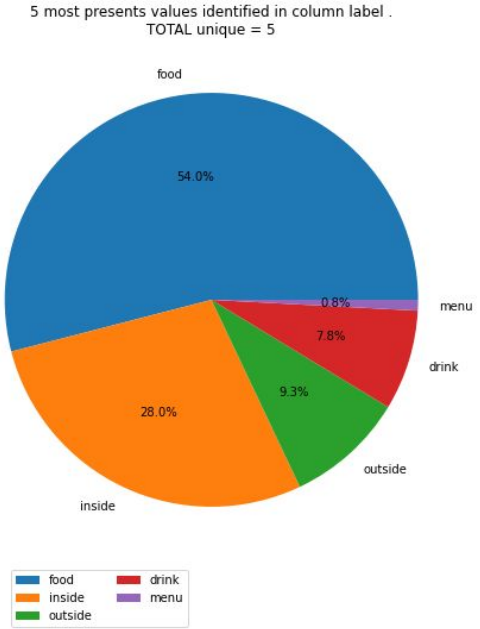
Live Demo

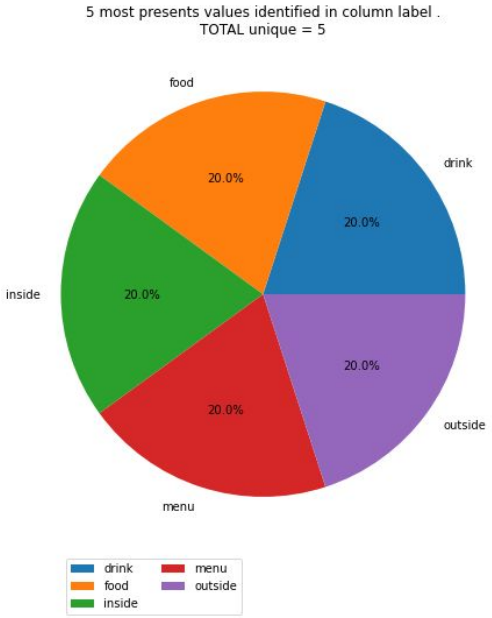
Back to AGENDA

Traitement des données Images

Sélection des Images / Photos

Sélection du jeu de donnée

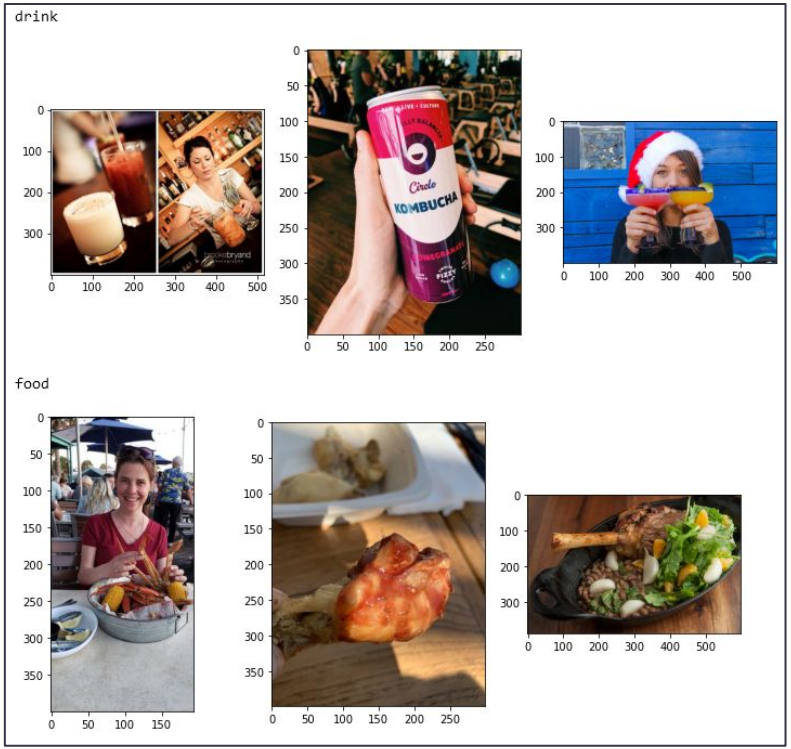


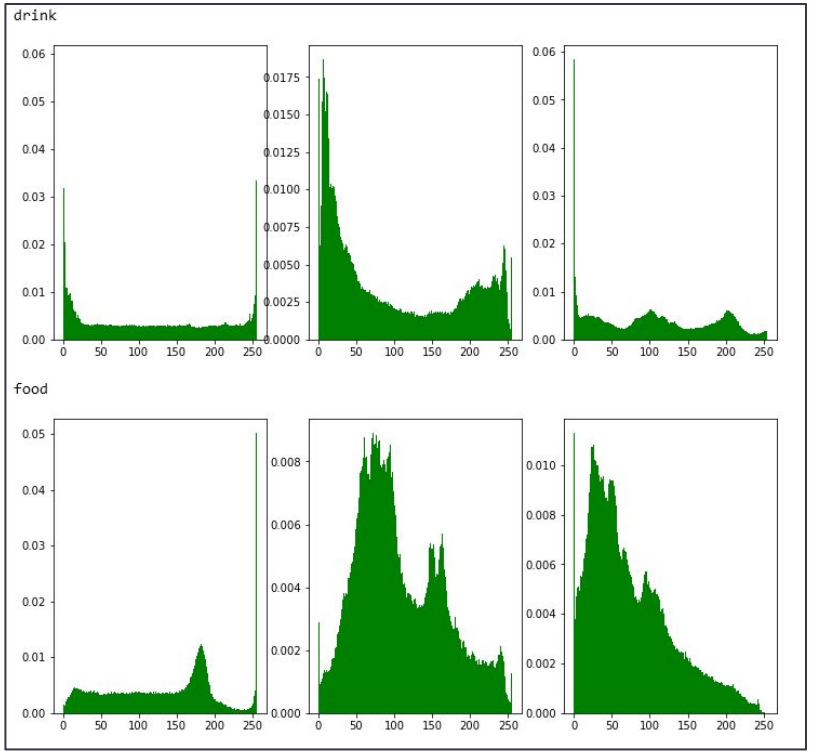


Suppression des images à taille nulle (ou égale à 1k octet)

Sélection par équilibre des catégories

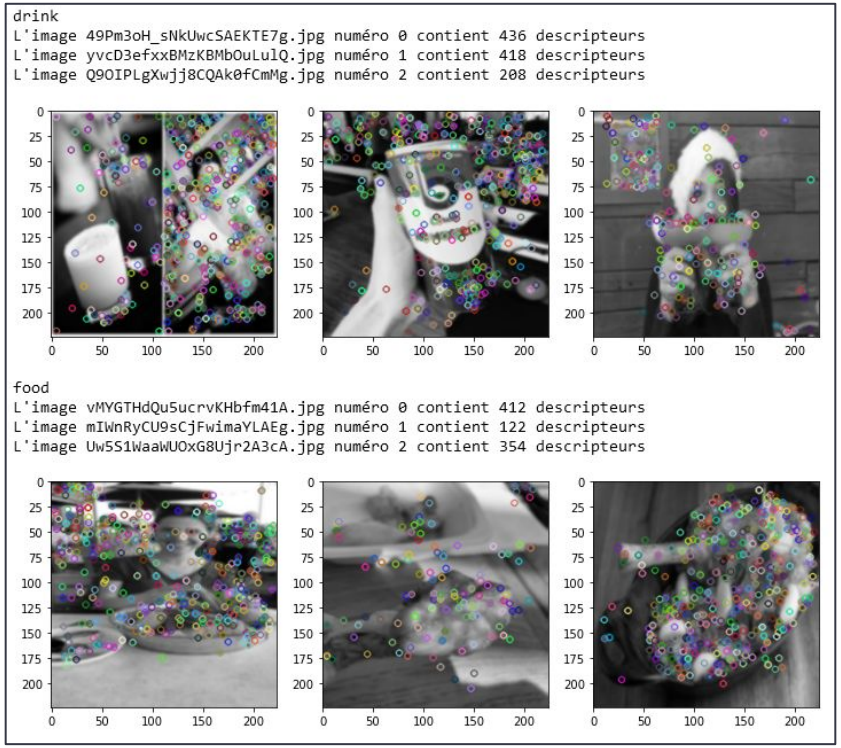
Photos sans preprocessing et Histogrammes





Preprocessing des Images et Détection des Descripteurs par SIFT

Preprocessing et Détecteur SIFT



Preprocess picture:

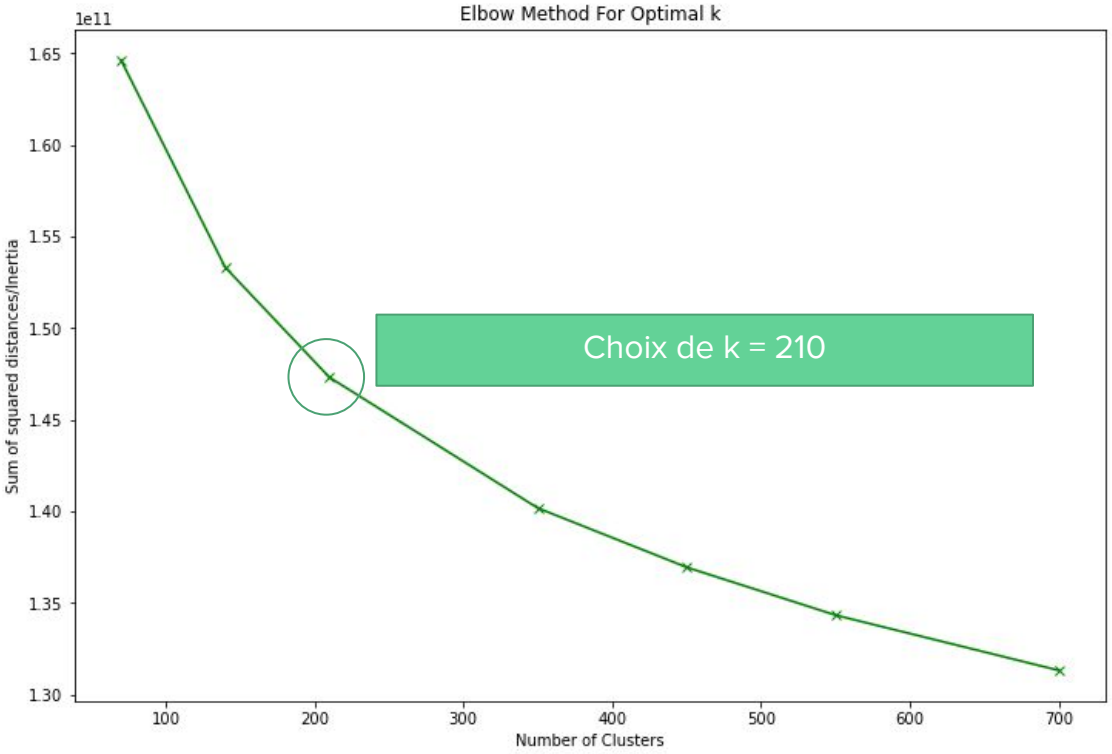
- Convert Colors (None, B&W, HSV)
- Resize (Yes, No)
- Equalize (None, Hist, Clahe)
- Blur (None, Normal ou Gaussian)

Detect Keypoints:

- SIFT Detect & Compute

Afficher Image et Keypoints

kMeans : Déterminer le nombre de descripteurs



MiniBatchKMeans avec k = variables

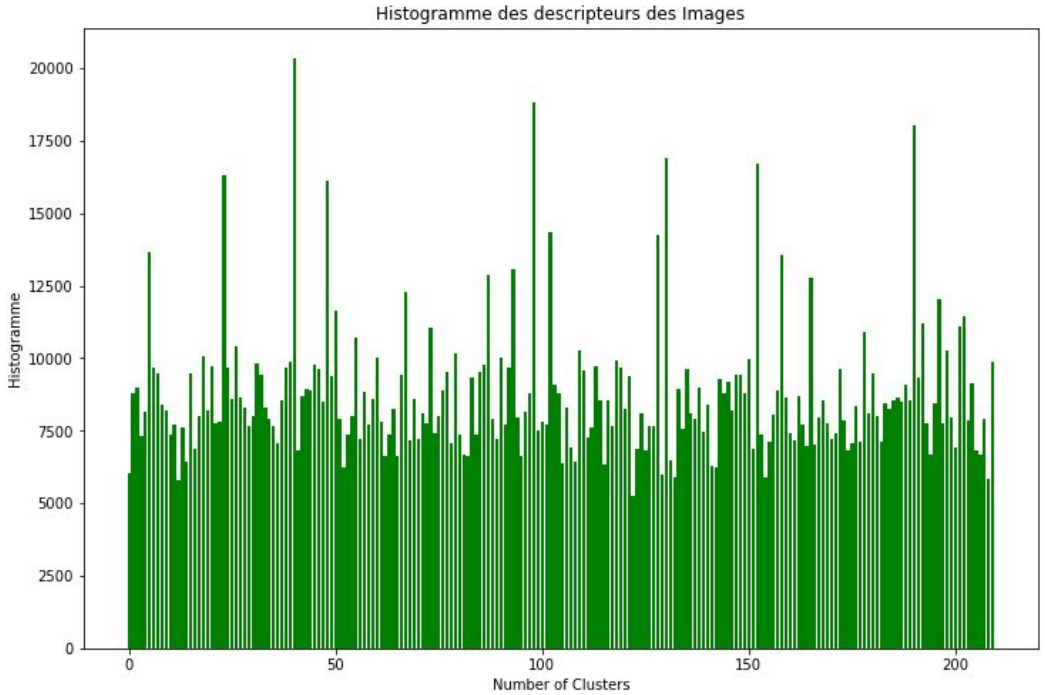
Détection du nombre de descripteurs par la méthode du coude.

Ce nombre de descripteurs sera ensuite notre dimension du BoW pour chacune des image.

Histogramme des descripteurs

Démarche:

1. predBoW = kmeans.predict sur les descripteurs
2. countBoW = collections.Counter sur predBoW
3. Somme des countBoW par descripteur



Création du Bag of Visual Words

	0	1	2	3	4	5	6	7	8	9	...	200	201	202	203	204	205	206	207	208	209
0	0	0	0	1	0	1	9	1	2	4	...	3	6	3	1	4	3	1	1	0	2
1	0	0	0	1	0	1	8	0	1	1	...	1	1	0	1	2	1	0	0	0	0
2	1	1	2	2	1	0	2	1	1	2	...	2	3	2	0	1	2	0	1	1	2
3	0	0	1	0	1	1	2	0	1	0	...	4	2	1	3	1	1	1	0	0	0
4	0	3	3	1	0	1	6	1	0	1	...	4	3	1	3	1	3	2	6	2	3
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4995	4	5	3	6	1	0	0	7	2	4	...	0	2	1	0	1	1	2	2	2	2
4996	0	0	0	0	0	4	0	0	0	1	...	0	1	3	0	1	0	0	0	1	1
4997	1	4	2	1	1	3	4	3	1	4	...	1	1	1	0	0	2	2	0	1	1
4998	1	2	1	1	4	6	1	4	4	2	...	4	0	1	6	2	2	0	1	2	0
4999	1	1	1	0	3	2	0	2	0	3	...	3	0	5	0	0	0	1	0	0	0

5000 rows × 210 columns

Bag of Visual Words Shape = 210 Descripteurs pour chacune de 5000 images

Le BoW est un catalogue de "mots visuels" qui décrivent les images.

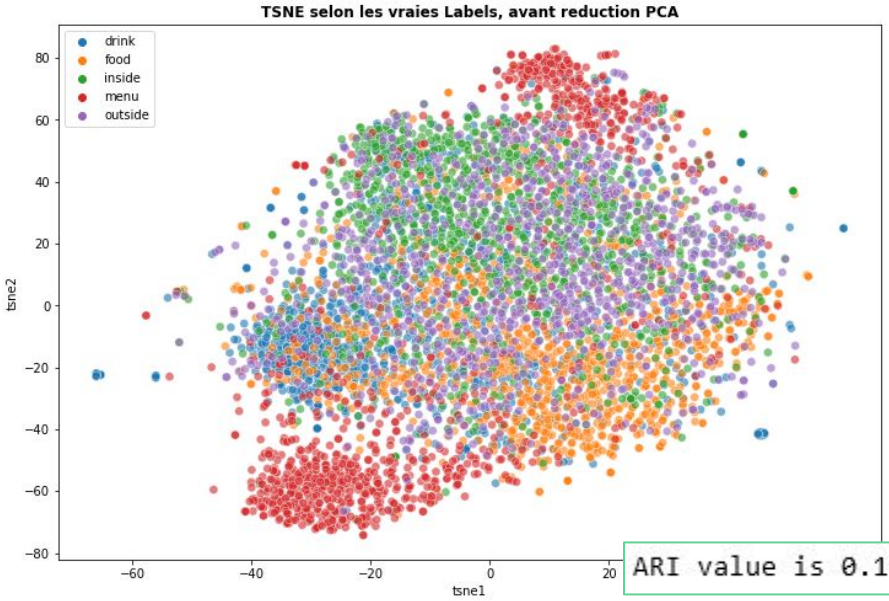
Nous réalisons un clustering sur les features / descripteurs que l'on vient d'extraire, par l'utilisation d'une segmentation kMeans (MiniBatchKMeans).

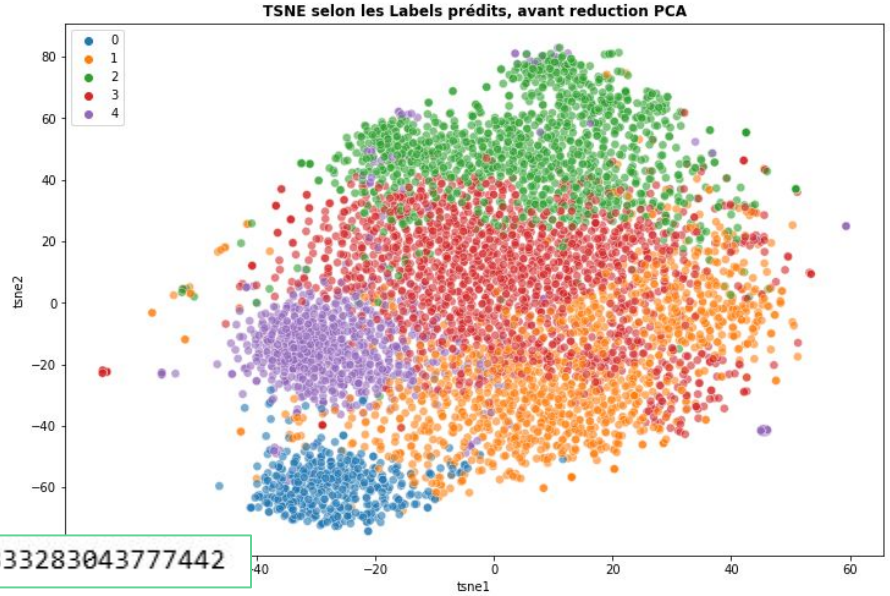
Ensuite on va compter le nombre d'occurence de chaques classes (= le Bag of Visual Words)

Classification non-supervisées

Visualisation T-SNE avant réduction de dimension

Prédire la classe de l'image (à partir des 210 features extraites via SIFT)





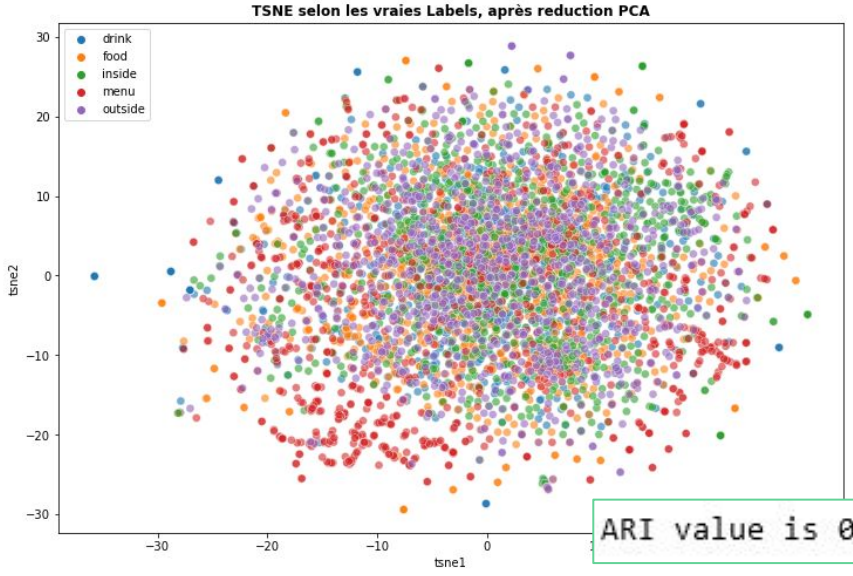

```
ARI value is 0.12333283043777442
```

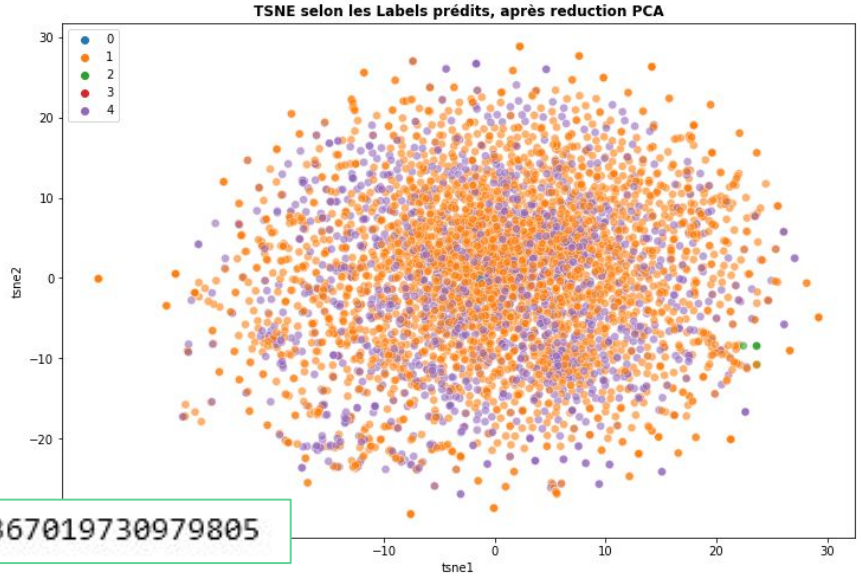
5 Classes représentées, mais segmentation floue (et ARI faible)

Visualisation T-SNE après réduction de dimension

Prédire la classe de l'image (à partir des 150 features réduites par ACP)

```
Explained variance after dimension reduction to 150 is = 0.9024
```



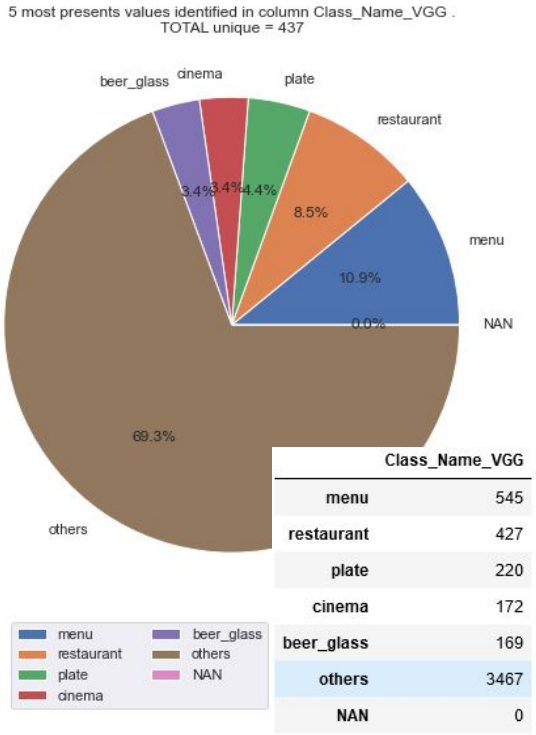



```
ARI value is 0.0052367019730979805
```

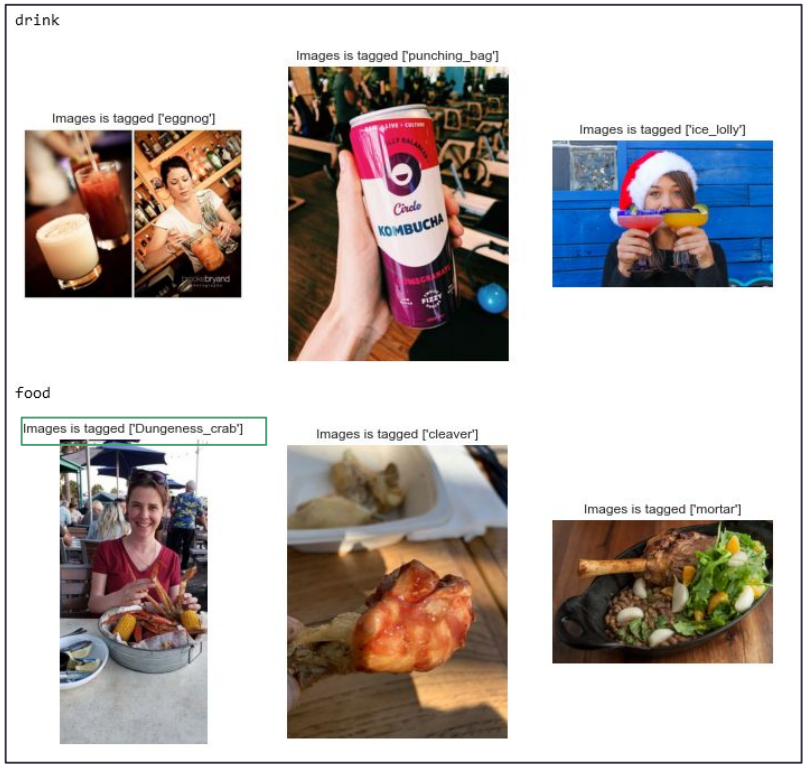
5 Classes non représentées = Inexploitable

Extraction de features avec un CNN pré-entraîné

Classification VGG Imagenet



Class_Name_VGG	
menu	545
restaurant	427
plate	220
cinema	172
beer_glass	169
others	3467
NAN	0



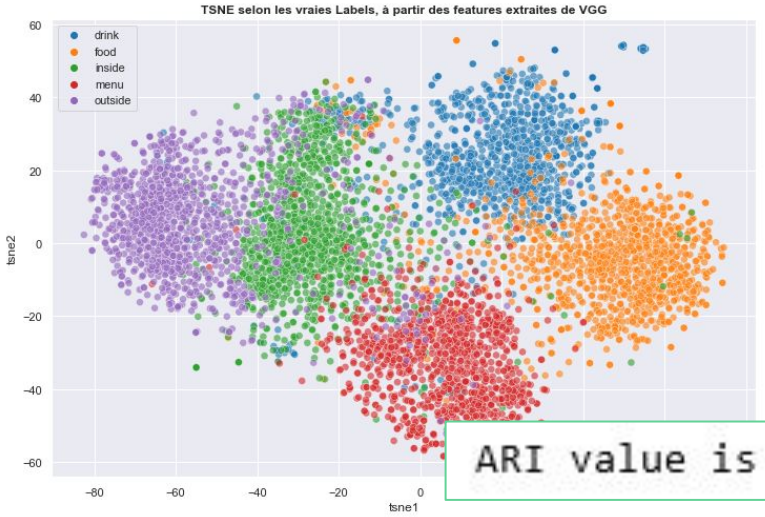
Prédiction des classes d'image (parmi les 1000 d'ImageNet).

Ne correspond pas au besoin métier

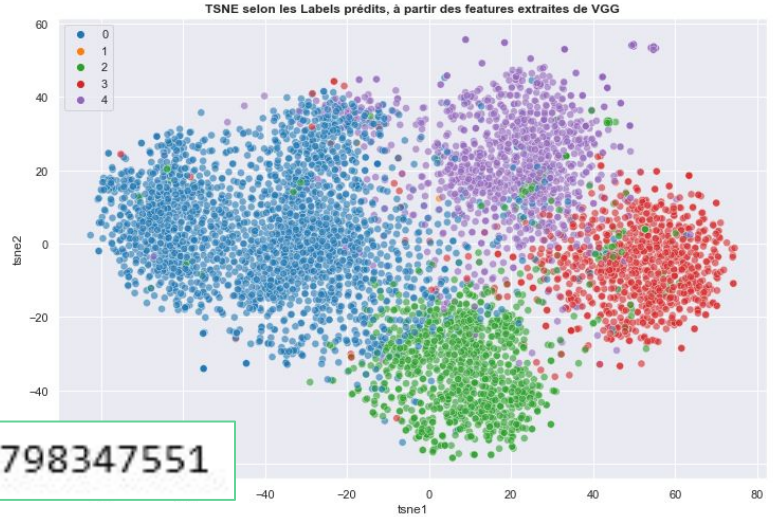
```
inside
outside
drink
food
menu
```

Visualisation T-SNE avant réduction de dimension

Prédire la classe de l'image (à partir des 512 features extraites du VGG pré-entrainé)



5000
Images



ARI value is 0.5034013798347551

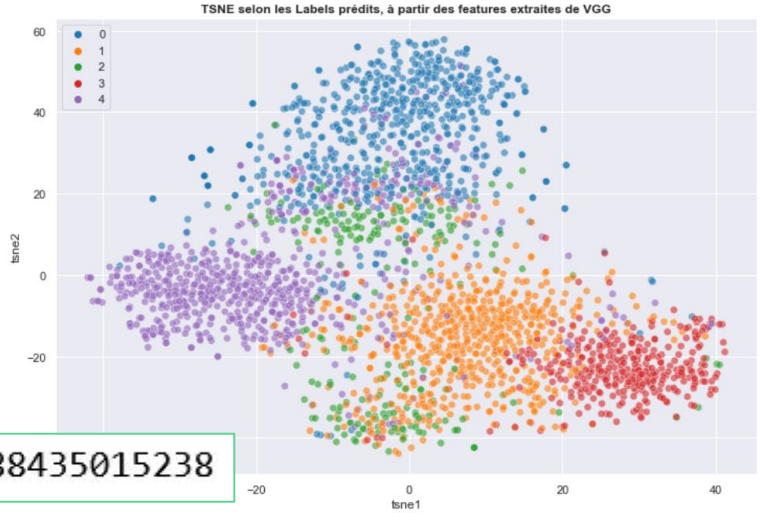
5 Classes non représentées = 4 uniquement !
La classe “inside” semble avoir disparue, et confondue avec “outside”
A évaluer avec le responsable métier ?

Visualisation T-SNE avant réduction de dimension

Prédire la classe de l'image (à partir des 512 features extraites du VGG pré-entrainé)

3750

Images

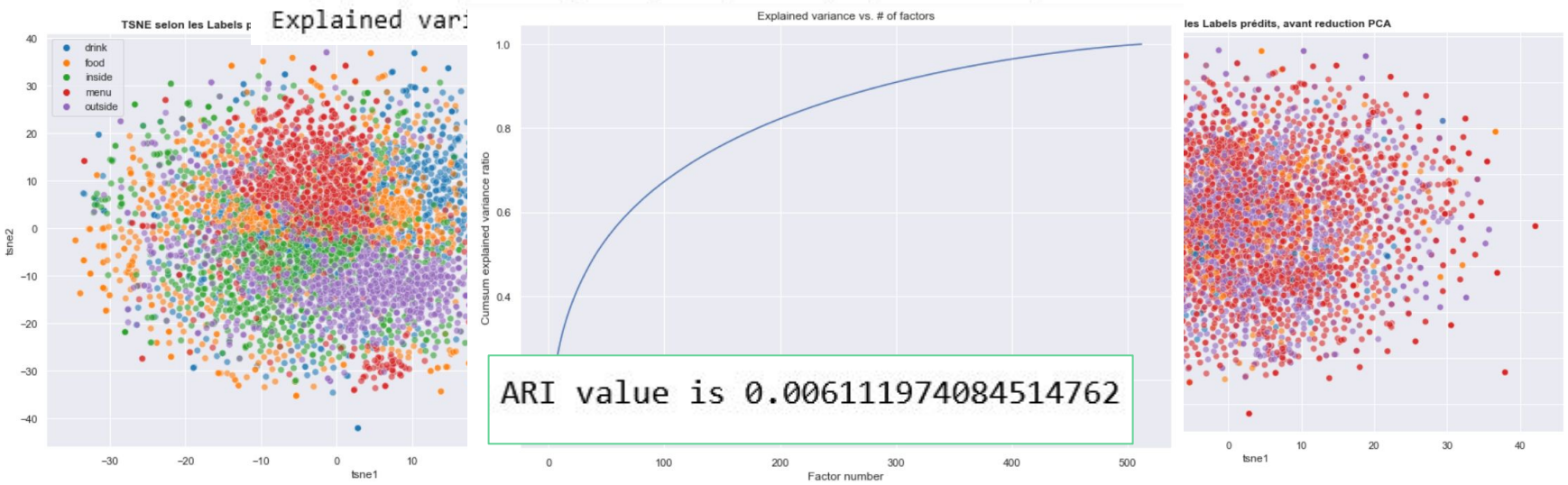


ARI value is 0.4499588435015238

5 Classes représentées ; Mais ARI plus faible !
La classe “inside” semble faire partie des prédictions.

Visualisation T-SNE après réduction de dimension

Prédire la classe de l'image (à partir des 300 features réduites par ACP)



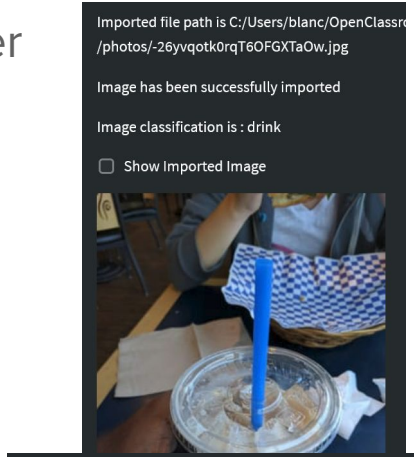
Explained variance after dimension reduction to 300 is = 0.9100000262260437

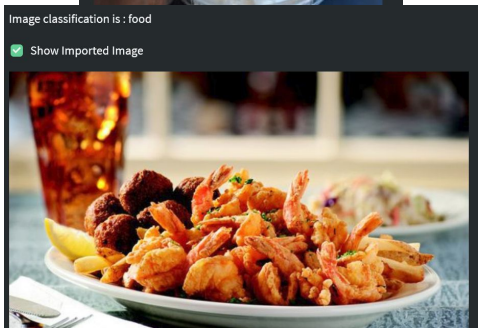
5 Classes non représentées = Inexploitable

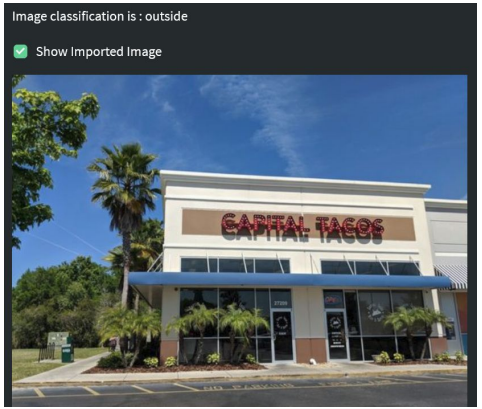
Traitement et Classification d’images

Utilisation de Streamlit pour générer une page HTML interactive.

1. Choix de l’image à importer
2. Traitement Image via pipeline VGG
3. Classification kmeans







Conclusion de l’ étude de faisabilité

Conclusions

Nous avons étudié 2 types de données : text review + images. Provenance du dataset Yelp (incluant les données de l'API)

La réduction de dimension ACP, nous permet de simplifier nos modèles. Les "Bag of Words" nous permettent de normaliser les textes et images, avant classification.

La classification kMeans démontre des possibilités de classer les données de manière non supervisée (Fiabilité < 50%)

Les techniques de réduction de dimensions (PyLDAVis + T-SNE) ont permis de représenter un ensemble de points d'un espace à grande dimension dans un espace de deux dimensions.